Model selection for simplicial approximation

Claire Caillerie, Bertrand Michel

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Model selection for simplicial approximation

Claire Caillerie — Bertrand Michel

N° 6981

Juillet 2009

Thème SYM

ISSN 0249-6399 ISRN INRIA/RR--6981--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

Model selection for simplicial approximation

Claire Caillerie*, Bertrand Michel*

Thème SYM — Systèmes symboliques
Équipe-Projet Geometrica

Rapport de recherche n° 6981 — Juillet 2009 — 25 pages

Abstract: In the computational geometry field, simplicial complexes have been used to describe an underlying geometric shape knowing a point cloud sampled on it. In this article, an adequate statistical framework is first proposed for the choice of a simplicial complex among a parametrized family. A least squares penalized criterion is introduced to choose a complex, and a model selection theorem states how to select the "best" model, with a statistical point of view. This result gives the shape of the penalty, and next, the so called "slope heuristics method" is used to calibrate the penalty from the data. Some experimental studies on simulated and real dataset illustrate the method for the selection of graphs in two dimensions.

Key-words: computational geometry, geometrical inference, simplicial complexes, model selection, penalization, slope heuristics.

* INRIA Saclay

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

Sélection de modèle pour l'approximation simpliciale

Résumé : Les complexes simpliciaux sont utilisés en géométrie algorithmique pour décire une forme géométrique à partir de points d'observation échantillonnés sur celle-ci. Cet article propose tout d'abord un cadre statistique adapté à la question du choix d'un complexe simplicial parmis une famille donnée. Un critère de moindres carrés est défini pour choisir un complexe simplicial, et un résultat de sélection de modèle établit comment choisir le "meilleur" complexe de la collection, selon un point de vue statistique. Ce résultat fournit la forme de la pénalité et la méthode dite de "l'heuristique de pente" permet dans un second temps de calibrer la pénalité à partir des données. Une étude expérimentale basée sur des données simulées et réelles illustrent l'utilisation de la méthode pour la sélection de graphes en dimension 2.

Mots-clés : géométrie algorithmique, inférence géométrique, complexes simpliciaux, sélection de modèles, pénalisation, heuristique de pente.

1 Introduction

Many methods have been proposed in statistics, data analysis or machine learning to extract information from a given dataset. The simplest and natural way to study a point cloud in $\mathbb{R}^D$ is the well known PCA (Principal Component Analysis). It consists of finding a linear subspace of $\mathbb{R}^D$ which preserves as much as possible the variance of the original dataset. If the data is actually located in the neighbourhood of a linear subspace, this elementary method provides an efficient representation of the data in a lower dimension space. During the nineties, some efforts have been made to adapt PCA to non linear situations, typically for point sets sampled on manifolds. For instance, "principal curves" methods [34, 21] have been defined to this aim. By definition, principal curves pass through the "middle" of a distribution, and they play the same role as the linear subspaces for PCA. Another solution is proposed in [35] by considering a combination of local linear PCA projections.

In the computational geometry community, the analysis of point clouds is also a popular research field. In this context, simplicial complexes have shown to be appropriate tools to describe underlying geometric shapes knowing a point cloud sampled on it. A simplicial complex is a collection of simplices such that any two simplices of the collection intersect along a common face if at all, and all the faces of a simplex of the collection belongs to the complex too. One interest in using simplicial complexes is that they allow not only dimension estimation procedures [e.g. 14] but also topological inference methods. Indeed, for a point cloud sampled on a geometrical object, some theoretical results show that the topology of a simplicial complex defined on the point cloud is the same as the topology of the original object, under some particular hypotheses [e.g. 13, 31]. Furthermore, effective algorithms for computing topology properties of simplicial complexes exist [38, 17], which makes possible the inference of topological properties in practice. Note that simplicial complexes have also been used for reconstruction [e.g. 9].

The typical and practical situation considered by this paper can be summarized as follows. Given a set of observed points $\mathcal{X} = \{X_1, \ldots, X_n\}$, a set of points called landmarks is defined from $\mathcal{X}$, or even directly extracted from $\mathcal{X}$. On this bunch of landmarks points, simplicial complexes can be defined according to a fixed rule chosen by the user. In many situations, this provides a large collection $(\mathcal{C}_\alpha)_{\alpha\in\mathcal{A}}$ of nested simplicial complexes indexed by a scale parameter α. Roughly speaking, a simplicial complex with a large α corresponds to a fine meshing. Each simplicial complex in the collection gives some particular geometric information on the data points, depending on the scale parameter α chosen (see for instance [12, 11, 31] for reconstruction and [13] for topological estimation). The choice of this scale parameter is thus of first importance since it will determine the geometric analysis of $\mathcal{X}$. A few contributions (see for instance[11]) have been proposed on how to choose a "convenient" scale paramater α, but as far as we know, there are no completely automatic "data-driven" criterion to do this selection. In any case the meaning of "convenient" has to be clarified. Most of the previously cited methods are determinist in the sense that they suppose that the observation points have been sampled exactly on an unknown geometric object and thus they are quite sensitive to outliers. An alternative approach is to consider this problem with a statistical point of view, which allows us to give a rigorous definition of what is the "convenient" scale paramater we would like to choose ideally. The aim of this article is first to propose an adequate statistical framework for the choice of a simplicial complex among a parametrized family, and second to give some mathematical answers to the problem of the complex choice. Our contribution should be considered as a first attempt to introduce model selection arguments [see e.g. 10, 28] into the realm of computational geometry using theoretical results of this statistical field.

Finding a general statistical model for the approximation of an unknown geometric object from noisy data with simplicial complexes is a difficult task. In the last years, a few pieces of work have been proposed in this direction. An interesting attempt is proposed in [3] where a Gaussian distribution is convoluted along a Delaunay graph. Thanks to an EM algorithm[16], the parameters of the complete model can be estimated. Unfortunately, this model is concerned with the case of graphs, and its mixture structure makes difficult its use for large complexes. A second attempt is made in [30]: for a point x on a submanifold $\mathcal{M}$ in $\mathbb{R}^N$ and a point y on the normal space

of $\mathcal{M}$ on x, the probability density measure $P(x, y)$ can be decomposed into $P(x, y) = P(x)P(y|x)$, where the marginal $P(x)$ is supported on the manifold $\mathcal{M}$ while $P(y|x)$ is a Gaussian distribution in the normal direction of $\mathcal{M}$ on x. Under these models and under a curvature condition on $\mathcal{M}$, the authors show that the topology of $\mathcal{M}$ can be learned from samples with high probability. An algorithm is also provided but it seems tricky to compute in practice since it relies upon the computations of so-called Cech complexes. Our statistical model embraces a different point of view. Since estimating a probability measure supported on a unknown geometrical object even in a parametric family is a tricky problem, we prefer to find a simplicial complex that correctly fits the observations rather than trying to figure out how the "true locations" of the observations have been sampled.

In this paper, the problem of choosing a simplicial complex is seen as a model selection problem in the context of density estimation. We only give here the main ideas of our model selection approach. The true density of the observations is unknown. Each simplicial complex $\mathcal{C}$ is associated to a set $S_{\mathcal{C}}$ of possible densities which models the fact that the observations are located in the neighbourhood of $\mathcal{C}$. The models are explicitly defined in Section 2. For each of these densities sets, a least squares estimator (LSE) can be defined, and the selection of a simplicial complex actually corresponds to the selection of a LSE. Roughly speaking, a simplicial complex with only a few components will not allow to approximate accurately the true density of the data. On the contrary, a simplicial complex which connects many nearby landmarks and which contains many simplexes will overfit the point cloud. Thus, a trade-off between the bias and the variance of the LSE has to be realized, which can be done by model selection procedures. A popular method for model selection is penalization. In our framework it consists in penalizing the least squares criterion by a penalty term that depends on the model complexity. The principle of selecting a model by penalizing loglikelihood or least squares criteria has emerged during the seventies. Akaike [1] proposed the AIC criterion (Akaike's information criterion) and Schwarz [33] suggested the BIC (Bayesian Information Criterion). BIC has already been used to select the numbers of vertices of a graph [20]. Note that these two classical criteria assume implicitly that the true distribution belongs to the model collection (see for instance [10]). Although the properties of these asymptotic criteria were tested in practice, little has been proved theoretically on this topic. A non asymptotic approach for model selection via penalization has emerged during the last ten years, mainly with works of Birgé and Massart [5] and Barron et al. [4] (an overview is available in [28]). In these works, the belonging of the true density to the model collection is not required. The aim of this approach is to define penalized data-driven criteria that lead to so-called oracle inequalities. The penalty function depends on the "complexity" of each model and also on the variety of the whole model collection. This approach has been carried out in several frameworks where penalty functions are explicitly assessed. In this paper, a general Gaussian model selection theorem is used to obtain a penalized criterion on a given family of simplicial complexes. Our result for simplicial complexes is quite general since it makes few assumptions on the complex family. The main advantage of this result is to provide the form of the penalty that should be used in practice. For instance, in the case of complexes of dimension one (graphs), it says that the penalty has to be chosen proportional to the logarithm of the graph length.

Nevertheless, our theorem for simplicial complex selection cannot be applied directly since the provided penalty is only known up to a multiplicative constant. A "slope heuristics" has been proposed in [6] to calibrate penalties when the penalty shape is known. The models we deal with are far from the situations for which theoretical results on the slope heuristics have been proved [6, 2]. Nevertheless, for some particular families of complexes, our simulations show that a "slope behaviour" for the LSE criterion can be observed. Furthermore, for a penalty calibrated by this method, the selected simplicial complex is closed to the "best" one, from a statistical point of view presented further. The simulations presented in this paper deal with graphs, they aim at illustrating the application of our method in some simple situations. Applications to more elaborate scenarios in higher dimension will be studied in forthcoming work.

Section 2 is devoted to the geometrical models and Section 3 is about model selection for simplicial complexes. Practical issues including the slope heuristics method are discussed in Section

4. The complete method is then carried out on simulated and real datasets in Section 5. A discussion section is finally given at the end of the paper.

2 Statistical models for geometry

In the sequel, for all $Q \in \mathbb{N}^*$, the space $\mathbb{R}^Q$ is equipped with the following normalized scalar product :

$$\forall u, v \in \mathbb{R}^Q, \quad \langle u, v \rangle := \frac{1}{Q} \sum_{i=1}^{Q} u_i v_i \ , \tag{1}$$

and the associated norm is denoted $\| \cdot \|$.

Suppose that we observe some points $X_1, \ldots, X_n$ located in the neighborhood of an unknown geometric object $\mathcal{G}$ embedded in $\mathbb{R}^D$. Generally speaking, our objective is to approximate $\mathcal{G}$ from the data points. Simplicial complexes have shown to be appropriate tools to describe underlying geometric shapes knowing a point cloud sampled on it. In consequence, we use these simplicial complexes to define a collection of possible estimators in order to infer $\mathcal{G}$. These ideas are now rigorously formalized by first introducing the following model.

Let $X_1, \ldots, X_n$ be some observed points such that

$$\forall i = 1, \ldots, n, \quad X_i = \bar{x}_i + \sigma \xi_i \quad \text{with} \quad \bar{x}_i \in \mathcal{G}, \tag{2}$$

where the original points $\bar{x}_i$ are unknown. The random variables ξ_i are independent standard Gaussian vectors of $\mathbb{R}^D$ and σ is the noise level. Let $\mathbf{X} = (X_1^t, \ldots, X_n^t)^t$ be the vector of length nD containing all the observations X_i . We also define $\bar{\mathbf{x}}$ and $\boldsymbol{\xi}$ in the same way. In the following, it will be convenient to consider the next equivalent statement of (2) in the space $\mathbb{R}^{nD}$:

$$\mathbf{X} = \bar{\mathbf{x}} + \sigma \boldsymbol{\xi} \quad \text{with} \quad \bar{\mathbf{x}} \in \mathcal{G}^n, \tag{3}$$

where $\boldsymbol{\xi}$ is a standard Gaussian vector of $\mathbb{R}^{nD}$. As it was said before, simplicial complexes provide some good approximations of an unknown geometric object. The best approximating point of $\bar{\mathbf{x}}$ belonging to $\mathcal{C}$ minimizes the quantity $\mathbf{t} \mapsto \|\mathbf{t} - \bar{\mathbf{x}}\|$. The least square estimator (LSE) of $\bar{\mathbf{x}}$ associated to the complex $\mathcal{C}$ is then defined by

$$\hat{\mathbf{x}}_{\mathcal{C}} := \text{argmin}_{\mathbf{t} \in \mathcal{C}^n} \|\mathbf{X} - \mathbf{t}\|^2. \tag{4}$$

In real situations, we are not dealing with a single complex: we need to choose one in a given collection on the knowledge of the vector of observations $\mathbf{X}$. Roughly speaking, a basic complex with only a few simplices will badly approximate $\mathcal{G}$ and the same is true for $\bar{\mathbf{x}}$, whereas a complex composed of too many simplices will tend to overfit the data. This facts exactly corresponds in statistics to the well known " bias-variance trade off " and it can be figured out by model selection methods.

This modeling can be related with the probabilist version of the well-known PCA method, see for instance [7], chap 12. Indeed, suppose that $\mathcal{G}$ is an unknown affine linear subspace of dimension p in $\mathbb{R}^D$. Then, the aim of PCA in this context is to find the subspace $\mathcal{V}$ of dimension p minimizing the quantity $\|\mathbf{X} - \hat{\mathbf{x}}_{\mathcal{V}}\|^2$ for the simple case where the variance matrix of ξ_i is the identity. In the case of a non linear object $\mathcal{G}$, we change the objective into finding a simplicial complex which efficiently fits the data, taking into account the overfitting phenomenon mentioned before. Note that in some cases, there is really a geometric structure and the noise in (2) models for instance a measurement error. But as a matter of fact, supposing that there is a "true" geometric structure $\mathcal{G}$ where the "true" points live is most of the time a mental construct.

Before presenting our model selection method for simplicial complex approximation, we first recall the definitions of well-known simplicial complex and we also give their main characteristics.

Simplicial complexes

A simplicial complex $\mathcal{C}$ is a set of simplices which satisfies the following conditions:

- Any face of a simplex from $\mathcal{C}$ is also in $\mathcal{C}$.
- The intersection of any two simplices s_1, $s_2 \in \mathcal{C}$ is either a face of both s_1 and s_2, or empty.

In the following, we call a k-simplex a simplex of dimension k. A simplicial complex is said to be k-homogeneous if each one of its simplices is either a k-simplex, or the face of a k-simplex of $\mathcal{C}$. In this paper, we are more interested in studying the support of a complex than its associated abstract space. By simplicial complexes we actually mean the support of the complexes by abusing the notation. We now give the main classical examples of simplicial complexes defined from a set of (landmark) points $Z \in \mathbb{R}^D$.

Abstract Complexes. A natural construction is the *Cech complex*: for all $\alpha > 0$, $\mathcal{C}^\alpha(Z)$ is the nerve of the open balls $\{B(z, \alpha) \ : \ z \in Z\}$ *ie.* a $p-$simplex $\sigma = [z_1 \dots z_p]$ belongs to $\mathcal{C}^\alpha(Z)$ if and only if the balls $\{B(z_j, \alpha) \ : \ j = 1 \dots p\}$ have non empty common intersection. Since balls are contractible in $\mathbb{R}^n$, Leray's Nerve Lemma implies that $\mathcal{C}^\alpha(Z)$ is homotopy equivalent to the union of balls [see 23, Corollary 4G3].

The *Rips complex* $\mathcal{R}^\alpha(Z)$ relaxes the Cech condition by allowing a simplex provided its vertices are pairwise within distance α *ie.* a $p-$simplex $\sigma = [z_1 \dots z_p]$ belongs to $\mathcal{R}^\alpha(Z)$ if and only if $\forall j, k \leq p | z_j - z_k| \leq \alpha$. The Rips complex is not homotopy equivalent to the Cech complex, but they are closely related [e.g. 13]. The Rips complex is much easier to compute than the Cech complex since it does not involve the computation of intersections of balls.

Geometric Complexes. For $p \in Z$, the Voronoi cell of p is the set of points in the ambient space closest to p than to other points of Z *ie.* $V(p) = \{x \in \mathbb{R}^d | d(x, p) \leq d(x, y), \forall p \in Z\}$. The Voronoi diagram decomposes $\mathbb{R}^D$ into convex cells. The *Delaunay complex* is the nerve of the Voronoi diagram *ie.* a $p-$simplex $\sigma = [z_1 \dots z_p]$ belongs to $\mathcal{D}\text{el}(Z)$ if and only if the Voronoi cells of z_i have non empty common intersection.

The *α-complex* is the nerve of the cover formed by the intersection of the union of α-balls (balls of radius α centered on the points of Z) and the Voronoi diagram[18]. This complex is homotopy equivalent to the union of balls of radius α [17], but is also embedded and has the same dimension as the ambient space. By construction, the α–complex is always a subcomplex of the Delaunay complex, so we may compute the former by computing the latter.

A last example is the *Witness Complex*, see [9] and [15] for details on its definition and its properties.

3 Model Selection on simplicial complexes

Choosing a simplicial complex to approximate an unknown geometrical object is not an easy question. In many situations, we deal with a collection of simplicial complexes indexed by a scale coefficient which needs to be calibrated. By proposing some LSE estimators (4) for each simplicial complex in a given collection, we recast the simplicial complex choice as a model selection problem. We first recall some general model selection results for the non linear Gaussian case.

3.1 Non linear Gaussian model selection

Suppose that we observe a random vector $\mathbf{X}$ in $\mathbb{R}^Q$ such that

$$\mathbf{X} = \bar{\mathbf{x}} + \sigma \boldsymbol{\xi} \tag{5}$$

where $\bar{\mathbf{x}}$ is an unknown vector of $\mathbb{R}^Q$, $\boldsymbol{\xi}$ is a standard Gaussian vector of $\mathbb{R}^Q$ and $\sigma > 0$ is the level noise. Let $(C_\alpha)_{\alpha \in \mathcal{A}}$ be a countable collection of compact sets in $\mathbb{R}^Q$ (a simplicial complex

for instance). The LSE of $\bar{\mathbf{x}}$ representing the model C_α is defined by

$$\hat{\mathbf{x}}_\alpha := \operatorname{argmin}_{\mathbf{t}\in C_\alpha} \|\mathbf{X}-\mathbf{t}\|^2.$$

In the sequel, the notation $\mathbb{E}_{\bar{\mathbf{x}}}$ denotes the expectation relative the probability law of $\mathbf{X}$ under the model (5), which can be parametrized by the determinist vector of true locations $\bar{\mathbf{x}}$. Note that $\hat{\mathbf{x}}_\alpha$ is a random variable which law depends on C_α an also on the distribution of $\mathbf{X}$.

In such a framework, a classical objective of model selection is the minimization of the estimation risk. The l^2 risk of the estimator $\hat{\mathbf{x}}_\alpha$ is then defined by

$$\mathcal{R}(\bar{\mathbf{x}},\alpha) = \mathbb{E}_{\bar{\mathbf{x}}}\left(\|\bar{\mathbf{x}}-\hat{\mathbf{x}}_\alpha\|^2\right).$$

Ideally, we would like to select the model

$$\alpha(\bar{\mathbf{x}}) = \operatorname{argmin}_{\alpha\in\mathcal{A}} \mathcal{R}(\bar{\mathbf{x}},\alpha).$$

Nevertheless, the model $\alpha(\bar{\mathbf{x}})$ and the quantity $\hat{\mathbf{x}}_{\alpha(\bar{\mathbf{x}})}$, called oracle, are unknown since they depend on the true value $\bar{\mathbf{x}}$. Actually, this oracle is a benchmark: A data-driven criterion has to be found to select an estimator such that its risk is close to the oracle risk.

At first sight, it seems natural to choose the estimator $\hat{\mathbf{x}}_\alpha$ of the collection that minimizes the quantity $\|\mathbf{X}-\hat{\mathbf{x}}_\alpha\|^2$. However, it is well known that such a procedure leads to select the largest models of the collection. Indeed, $\|\mathbf{X}-\hat{\mathbf{x}}_\alpha\|^2$ is not a consistent estimator of the risk of $\hat{\mathbf{x}}_\alpha$; the least squares term underestimates the risk of $\hat{\mathbf{x}}_\alpha$ by a term which is of the order of the "model complexity" [see for instance 22, chapter 7]. The principle of selecting a model by using a penalized criterion to avoid this overfitting phenomenon has emerged with the works of Akaike [1], Mallows [26] and Schwarz [33]. In our context, a model selection via penalization procedure consists of considering some proper penalty function $\text{pen} : \alpha \in \mathcal{A} \mapsto \text{pen}(\alpha) \in \mathbb{R}^+$ and of selecting $\hat{\alpha}$ minimizing the associated l^2 criterion

$$\text{crit}(\alpha) = \|\mathbf{X}-\hat{\mathbf{x}}_\alpha\|^2 + \text{pen}(\alpha).$$

The resulting selected estimator is denoted $\hat{\mathbf{x}}_{\hat{\alpha}}$. Obviously, the main difficulty of this approach is to choose a convenient penalty in order to select a estimator close to the oracle. For instance, the well known AIC penalty is $2d_\alpha\hat{\sigma}^2/n$ where $\hat{\sigma}^2$ is an estimator of the noise variance and d_α the "number of parameters" estimated by $\hat{\mathbf{x}}_{\hat{\alpha}}$. But what is the number of parameters of an estimator $\hat{\mathbf{x}}_{\mathcal{C}}$ associated to a simplicial complex $\mathcal{C}$ as defined in Section 2 ? This shows that the classical methods of penalization cannot be easily applied in our context.

A new theory of penalization with a non asymptotic approach has been developed in the nineties, with the works of Birgé and Massart[see 28] among others. The final purpose of a non asymptotic approach for model selection is to obtain a penalty function and an associated oracle inequality, allowing to compare the risk of the penalized LSE $\hat{\mathbf{x}}_{\hat{\alpha}}$ with the benchmark

$$\inf_{\alpha\in\mathcal{A}} \mathbb{E}_{\bar{\mathbf{x}}}\left(\|\bar{\mathbf{x}}-\hat{\mathbf{x}}_\alpha\|^2\right).$$

In [5], such a non asymptotic model selection result is obtained for collections of linear Gaussian models, namely for collections of linear subspaces C_α. In this case, they show that a good penalty has to be chosen proportional to the model dimension. Here, as for the geometric models defined in Section 2, the sets C_α are not supposed to be linear spaces. Section 4.4 in [28] shows that efficient penalties can be defined on nonlinear Gaussian models by using the metric entropy. Indeed, metric entropy allows to quantify the size of a metric space.

Let S be a set in the normed space $\left(\mathbb{R}^Q, \|\cdot\|\right)$ and $r>0$. A finite subset S_r of S with maximal cardinality such that for every distinct points x and y in S_r one has $\|x-y\|>r$, is a r-net of S. The maximum cardinality is denoted $\mathrm{N}\left(S,\|\cdot\|,r\right)$. The r-entropy of S is defined by

$$\mathrm{H}(S,\|\cdot\|,r) := \ln\ \mathrm{N}\left(S,\|\cdot\|,r\right).$$

It is generally easier to compute r-covering number : let $\mathrm{N}'(S, \|\cdot\|, r)$ be the minimal number of balls of radius r centred on points of S to cover S. Then, we have

$$\mathrm{N}'(S, \|\cdot\|, r) \leq \mathrm{N}(S, \|\cdot\|, r) \leq \mathrm{N}'(S, \|\cdot\|, r/2). \tag{6}$$

For all $\alpha \in \mathcal{A}$, the auxiliary entropic function Φ_α is defined by

$$\Phi_\alpha(u) = \kappa \int_0^u \sqrt{\mathrm{H}(C_\alpha, \|\cdot\|, r)}\, dr.$$

In the sequel, the constant κ can be taken greater or equal to 96 although this value is not optimal for the application of Theorem 1. For all $\alpha \in \mathcal{A}$ let d_α defined by the equation (if it exits)

$$\Phi_\alpha\left(\frac{2\sigma\sqrt{d_\alpha}}{\sqrt{Q}}\right) = \frac{\sigma d_\alpha}{\sqrt{Q}}. \tag{7}$$

Also suppose that some weights w_α fulfills

$$\sum_{\alpha \in \mathcal{A}} e^{-w_\alpha} = \Sigma < \infty. \tag{8}$$

Under the previous hypotheses, Theorem 4.18 in [28] can be rewritten as follows:

Theorem 1. *Let $\eta > 1$ and take*

$$\mathrm{pen}(\alpha) \geq \eta \frac{\sigma^2}{Q}\left(\sqrt{d_\alpha} + \sqrt{2w_\alpha}\right)^2. \tag{9}$$

Then, almost surely, there exists a minimizer $\hat{\alpha}$ of the penalized criterion

$$\mathrm{crit}(\alpha) = \|\mathbf{X} - \hat{\mathbf{x}}_\alpha\|^2 + \mathrm{pen}(\alpha).$$

Defining the penalized estimator by $\hat{\mathbf{x}}_{\hat{\alpha}}$, the following risk bound holds for all $\bar{\mathbf{x}} \in \mathbb{R}^Q$

$$\mathbb{E}_{\bar{\mathbf{x}}}\|\hat{\mathbf{x}}_{\hat{\alpha}} - \bar{\mathbf{x}}\|^2 \leq c_\eta \left[\inf_{\alpha \in \mathcal{A}} \left\{d(\bar{\mathbf{x}}, C_\alpha)^2 + \mathrm{pen}(\alpha)\right\} + \frac{\sigma^2}{Q}(\Sigma + 1)\right] \tag{10}$$

where c_η depends only on η and $d(\bar{\mathbf{x}}, C_\alpha) := \inf_{\mathbf{y} \in C_\alpha} \|\bar{\mathbf{x}} - \mathbf{y}\|$.

Several remarks can be given about this theorem. The weights w_α are introduced to control the richness (size) of the model collection ; this is the signification of Condition (8) which is used in the proof to control events on the whole model collection. These weights have to be large enough too fullfill Condition (8) and make Σ small in the risk bound (10). But they should not be too large since they are also involved in the penalty bound (9). This bound is also proportional to the quantity d_α which plays a similar role as the dimension in the case of linear models. Indeed; if C_α is a linear space of dimension d'_α, it can be easily shown that $d_\alpha = d'_\alpha$ [see 28, p130]. Thus, this result shows that a model has to be penalized by a quantity that is roughly speaking proportional to the metric dimension of the model.

Strictly speaking, the risk bound (10) is not exactly an oracle inequality since the risk of the selected estimator is not compared to the risks of all the estimators in the collection. An accurate comparison of $\mathbb{E}_{\bar{\mathbf{x}}}\|\hat{\mathbf{x}}_\alpha - \bar{\mathbf{x}}\|^2$ and $d(\bar{\mathbf{x}}, C_\alpha)^2 + \mathrm{pen}(\alpha)$ is possible in the linear Gaussian case (see [28] p.91). Generally speaking an rigorous oracle inequality is difficult to set since it requires to know the shape of the risks of all the estimators in the collection.

For applications of this theorem, most of the times the function Φ_α can be only bounded since it is usually impossible to compute the exact value of $\mathrm{H}(C_\alpha, \|\cdot\|, r)$. Fortunately, an upper bound of Φ_α is sufficient to propose a lower bound on the penalty from (9). The reader is referred to Chapter 4 in [28] for more details about Gaussian model selection. We now return to the particular case of the geometrical models defined in Section 2.

3.2 Main results

For a k-simplex s in $\mathbb{R}^D$, let Δ_s be the diameter of the smallest including ball of s for the normalized norm (1) in $\mathbb{R}^D$. Then, for a k-homogeneous simplicial complex $\mathcal{C}$ in $\mathbb{R}^D$, let $|\mathcal{C}|_k := (\sum_{s\in\mathcal{C}^+} \Delta_s^k)^{1/k}$ and $\delta_{\mathcal{C}} := \inf_{s\in\mathcal{C}^+} \Delta_s$ where $\mathcal{C}^+$ is the subset of simplices of $\mathcal{C}$ of maximal dimension k. We start with the following entropic result on simplicial complexes.

Proposition 1. *For all k-homogeneous simplicial complex $\mathcal{C}$ of $\mathbb{R}^D$ and all $r \leq \delta_{\mathcal{C}}$*

$$N(\mathcal{C}^n, \|\cdot\|, r) \leq \left(\frac{4|\mathcal{C}|_k}{r}\right)^{nk}.$$

Proof. Let s be a k-simplex of $\mathcal{C}$ and suppose witout loss of generality that $0 \in s$. Then, there exists a linear subspace F in $\mathbb{R}^D$ of dimension k such that $s \subset F$. Let Δ_s be the diameter of s: there exists a ball B_s of diameter Δ_s such that $s \subset B_s \cap F$. Using for instance [32], p.63, for all $r > 0$, there exists an r-covering of $B_s \cap F$ by a family $B_1 \cap F, \ldots, B_N \cap F$ with $B_i = B(u_i, r)$, $u_i \in F$ and where

$$N := N'(s, \|\cdot\|, r) \leq \left(1 + \frac{\Delta_s}{r}\right)^k.$$

Let $s_1, \ldots, s_{\mathrm{L}}$ be the family of k-simplices of $\mathcal{C}$, then for all $r \leq \delta_{\mathcal{C}}$

$$\begin{aligned} N'(\mathcal{C}, \|\cdot\|, r) &\leq \sum_{i=1}^{L} N'(s_i, \|\cdot\|, r) \\ &\leq \sum_{i=1}^{L} \left(1 + \frac{\Delta_{s_i}}{r}\right)^k \\ &\leq \left(2\frac{|\mathcal{C}|_k}{r}\right)^k \end{aligned}$$

since $\Delta_{s_i} \geq r$ for all i. Let $\mathcal{U}$ be the family of centers corresponding to such a r-covering of $\mathcal{C}$, with the cardinal of $\mathcal{U}$ less than $\left(2\frac{|\mathcal{C}|_k}{r}\right)^k$. For all $\mathbf{u} = (u_1^t, \ldots, u_n^t)^t \in \mathbb{R}^{nD}$, since $\|\mathbf{u}\|^2 = \frac{1}{n}\sum_{i=1}^n \|u_i\|^2$, then we have $\prod_{i=1}^n B(u_i, r) \subset B(\mathbf{u}, r)$. Note that in the last statements we use the same notations $\|\cdot\|$ for the normalized norms (1) in $\mathbb{R}^D$ or $\mathbb{R}^{nD}$. Thus, the family of balls $B(\mathbf{u}, r)$, where the u_i are chosen in $\mathcal{U}$, covers $\mathcal{C}^n$. Finally,

$$N'(\mathcal{C}^n, \|\cdot\|, r) \leq \left(\frac{2\ |\mathcal{C}|_k}{r}\right)^{nk}$$

and with (6):

$$N(\mathcal{C}^n, \|\cdot\|, r) \leq \left(\frac{4\ |\mathcal{C}|_k}{r}\right)^{nk}.$$

□

We are now in position to state a model selection result for simplicial complexes. Let $(\mathcal{C}_\alpha)_{\alpha\in\mathcal{A}}$ be a given collection of k-homogeneous simplicial complexes in $\mathbb{R}^D$. Suppose that there exists some weights w_α such that

$$\sum_{\alpha\in\mathcal{A}} e^{-w_\alpha} = \Sigma < \infty.$$

Let $\mathbf{X}$ be the observation vector with the distribution defined by (3). Proposition 1 allows us to apply Theorem 1 with the models $C_\alpha = \mathcal{C}_\alpha^n$ and $Q = nD$. For each $i = 1, \ldots, n$, let $\hat{x}_{\alpha i}$ be the closest point of X_i belonging to $\mathcal{C}_\alpha$. Thus, $\hat{\mathbf{x}}_\alpha = (\hat{x}_{\alpha 1}^t, \ldots, \hat{x}_{\alpha n}^t)^t$ is the least squares estimator of $\bar{\mathbf{x}}$ associated to the model $\mathcal{C}_\alpha^n$.

Theorem 2. *Under the previous hypotheses, also suppose that for all* $\alpha \in \mathcal{A}$*,*

$$\sigma \leq \delta_{\mathcal{C}_\alpha} \sqrt{\frac{D}{k}} \left[4\kappa \left(\sqrt{\ln \frac{4|\mathcal{C}_\alpha|_k}{\delta_{\mathcal{C}_\alpha}}} + \sqrt{\pi} \right) \right]^{-1} . \tag{11}$$

There exists some absolute constants c_1 *and* c_2 *such that for all* $\eta > 1$*, if*

$$\mathrm{pen}(\alpha) \geq \eta \frac{\sigma^2}{Q} \left(c_1 nk \left[\ln \frac{|\mathcal{C}_\alpha|_k \sqrt{D}}{\sigma \sqrt{k}} + c_2 \right] + 4 w_\alpha \right), \tag{12}$$

then, almost surely, there exists a minimizer $\hat{\alpha}$ *of the penalized criterion*

$$\mathrm{crit}(\alpha) = \|\mathbf{X} - \hat{\mathbf{x}}_\alpha\|^2 + \mathrm{pen}(\alpha)$$

and the penalized estimator $\hat{\mathbf{x}}_{\hat{\alpha}}$ *satisfies the following risk bound*

$$\mathbb{E}_{\bar{\mathbf{x}}} \|\hat{\mathbf{x}}_{\hat{\alpha}} - \bar{\mathbf{x}}\|^2 \leq c_\eta \left[\inf_{\alpha \in \mathcal{A}} \left\{ d(\bar{\mathbf{x}}, \mathcal{C}_\alpha^n)^2 + \mathrm{pen}(\alpha) \right\} + \frac{\sigma^2}{Q} (\Sigma + 1) \right]. \tag{13}$$

Condition (11) means that the complexes in the collection should not contain any k-simplexes with a diameter of the order of the noise level σ. This is natural since it would not be relevant to fit some simplices of this small scale on the observed data. This also means that the landmarks used to define the complexes should not be chosen too close of each other.

Several remarks can be given about the constants involved in this result. First, note that Condition (11) implies that the logarithm term in the penalty is always positive. Next, the constant c_η is the same as in Theorem 1 and it only depends on η. The proof shows that we can choose $c_1 \leq 16\kappa^2$ and $c_2 \leq \pi + \ln \frac{1}{\kappa\sqrt{\pi}}$. Nevertheless, these bounds have no interest since they are surely far from being optimal. These remarks about the constants suggests that this theorem has to be considered from a qualitative point of view : the main contribution of this result is to give the penalty shape, although it does not directly provide a penalty function usable for the practice.

The shape of the penalty function in our case is quite different than penalty shapes used in previous model selection works in the spirit of the results initiated by Birgé and Massart. Indeed, for instance for linear Gaussian model, for density estimation or for regression, the penalty is generally taken proportional to number of free parameters, see [28] for an overview. In our context, the number of free parameters nk is constant over the collection of simplicial complexes and thus the relevant term in the penalty bound (12) is the "size measurement" $\ln |\mathcal{C}_\alpha|_k$ of the complex. We will see that in spite of the logarithm term, this quantity varies a lot over the collection in practice. The penalty also depends on the weighs w_α. By analogy with the case of linear models [see 28, p.91], we can choose weighs such that $w_\alpha = L \ln |\mathcal{C}_\alpha|_k$ with

$$\sum_{\alpha \in \mathcal{A}} \frac{1}{x_\alpha^L} = \Sigma < \infty$$

where $L > 0$. Then, the lower bound (12) is actually proportional to $\ln |\mathcal{C}_\alpha|_k$. The application section 5 focuses on the study of graphs ($k = 1$). In this particular case, the term $\ln |\mathcal{C}_\alpha|_k$ exactly corresponds to the logarithm of the graph length, which is easy to compute. In practice, some additional work is next necessary to efficiently calibrate a penalty of this form. This problem is tackled in Section 4 with the "slope heuristics" method.

According to (3), the true positions $\bar{x}_1, \ldots, \bar{x}_n$ are located on a geometric object $\mathcal{G}$ of dimension $k < D$, but we did no hypotheses on how the $\bar{x}_i$ are sampled on $\mathcal{G}$. An integrated version of the risk bound (13) can be easily deduced, which will be usefull for the justification of the slope heuristics in Section 4. Let $\bar{x}$ be a random variable distributed according to a probability measure μ on $\mathcal{G}$. The risk bound (13) can be rewritten as follows ($n = 1$):

$$\mathbb{E}\left(\|\hat{x}_{\hat{\alpha}} - \bar{x}\|^2 \,|\, \bar{x} \right) \leq c_\eta \left[\inf_{\alpha \in \mathcal{A}} \left\{ d(\bar{x}, \mathcal{C}_\alpha)^2 + \mathrm{pen}(\alpha) \right\} + \frac{\sigma^2}{Q} (\Sigma + 1) \right] \quad \text{a.s..} \tag{14}$$

where the expectation is relative to the law of an observation $X = \bar{x} + \sigma\xi$ in $\mathbb{R}^D$. Note that the right side term in (14) is well defined since $\mathcal{C}_\alpha$ is a compact set and thus $\|\hat{x}_{\hat{\alpha}} - \bar{x}\|^2 < \infty$ μ-almost surely. We then define the integrated risk $\mathcal{R}(\mu, \alpha)$ on $\mathcal{G}$ by integrating the risk (14) on $\mathcal{G}$ according to the law of $\bar{X}$:

$$\begin{aligned} \mathcal{R}(\mu,\alpha) &:= \int_{\bar{x}\in\mathcal{G}} \mathbb{E}_{\bar{x}} \|\hat{x}_{\hat{\alpha}} - \bar{x}\|^2 \, d\mu(\bar{x}) \\ &\leq c_\eta \left[\inf_{\alpha\in\mathcal{A}} \left\{ \int_{\bar{x}\in\mathcal{G}} d(\bar{x}, \mathcal{C}_\alpha)^2 \, d\mu(\bar{x}) + \mathrm{pen}(\alpha) \right\} + \frac{\sigma^2}{Q}(\Sigma + 1) \right]. \end{aligned} \tag{15}$$

4 Penalty calibration by the slope heuristics

The aim of this section is to complete the theoretical results of last section in order to obtain an usable data-driven method for the selection of a simplicial complex in a given collection. Indeed, Theorem 2 does not directly provide an usable model selection criterion since the lower bound (12) is defined up to unknown constants.

Suppose that the considered collection of complexes is only composed of k-homogeneous complexes. Thus, for a fixed observed sample and a given simplicial complex collection, n, D and k can be seen as constants in (12). Theorem 2 and the remarks following it show that for the practice, penalties have to be chosen proportional to $\ln |\mathcal{C}_\alpha|_k$, if we only consider the principal term in the lower bound (12). Note that the constant c_2 could be also taken into account by using an elaborated penalty calibration as in [24]. To a first approximation, we deal with penalties proportional to $\ln |\mathcal{C}_\alpha|_k$ in the sequel of the paper.

Thus, the penalty shape is known, but some additional work is necessary to define a completely data-driven model selection criterion. A practical method called "*slope heuristics*", based on a mixture of theoretical and heuristics ideas for defining efficient penalty functions from the data, is proposed in [6]. This heuristics is proved only in the framework of Gaussian regression with a homoscedastic fixed design [6] and more recently generalized in the heteroscedastic random-design case [2]. Nevertheless applications of this method are developed in many other frameworks: For instance, in multiple change points detection [24], in genomic applications [37] and in Gaussian Markov random fields [36]. The application of the slope heuristics to Gaussian mixture models has also been studied in [29].

The collections we are intersted in for the applications are composed of simplicial complexes of the same "kind", for instance we do not mix α-complexes with α-Rips in a same collection. All the complexes presented in Section 2 can be parametrized by a real positive coefficient $\alpha(\mathcal{C})$ giving the "scale" of the complex. This is the case for instance for collections of α-complexes or α-Rips. Thus, the simplicial complexes are supposed to be indexed by their scale parameter: $\alpha = \alpha(\mathcal{C})$, and the model index set $\mathcal{A}$ is exactly the discrete subset of $\mathbb{R}^+$ of all the possible simplicial complex scales.

The slope heuristics method can be summarized as follows for the geometrical framework of this paper:

1. For each simplicial complex, compute the sum of squares $SS(\alpha) := \|\hat{\mathbf{x}}_\alpha - \mathbf{X}\|^2$.
2. Plot the point cloud $\{\ln |\mathcal{C}_\alpha|_k, SS(\alpha)\}_{\alpha\in\mathcal{A}}$ and check that a linear trend is observed for large α.
3. Compute the (negative) slope $\hat{\beta}$ of the linear regression of $SS(\alpha)$ on $\ln |\mathcal{C}_\alpha|_k$ for large α.
4. Select the simplicial complex in the collection minimizing

$$\mathrm{crit}(\alpha) = \|\bar{\mathbf{x}} - \hat{\mathbf{x}}_\alpha\|^2 - 2\hat{\beta} \ln |\mathcal{C}_\alpha|_k \ .$$

The previous description is sufficient for the reading of the rest of the paper but some additional justifications follows for a better understanding of the slope heuristics in our context.

Rationale for the slope heuristics

For the sequel, it will be convenient to suppose that the vectors of true positions $\bar{x}_i$ are sampled according to the probability measure μ on $\mathcal{G}$. Let $\bar{\mathbf{x}}_\alpha$ be the closest point of $\bar{\mathbf{x}}$ in $\mathcal{C}_\alpha^n$, and thus $\bar{\mathbf{x}}$ and $\bar{\mathbf{x}}_\alpha$ are both random variables.

The least squares method can be rewritten as follows. For $\mathbf{t} \in \mathbb{R}^{nD}$, let $\gamma(\mathbf{t}, y) := \|\mathbf{t}\|^2 - 2\langle \mathbf{t}, y\rangle$ and $\gamma_n(t) := \|\mathbf{t}\|^2 - 2\langle \mathbf{t}, \mathbf{X}\rangle$ be respectively the contrast and the empirical contrast associated to the least square procedure. Then, the LSE on $\mathcal{C}_\alpha^n$ is exactly

$$\hat{\mathbf{x}}_\alpha = \operatorname{argmin}_{\mathbf{t}\in C_\alpha^n} \gamma_n(\mathbf{t}).$$

whereas

$$\bar{\mathbf{x}}_\alpha = \operatorname{argmin}_{\mathbf{t}\in C_\alpha^n} \gamma(\mathbf{t}, \bar{\mathbf{x}}).$$

The introduction of these two contrasts γ and γ_n makes easier the description of the slope heuristics in the sequel. Let $\mathbf{X}'$ be a random variable in $\mathbb{R}^D$, which is independent of X and with the same law than X knowing $\bar{\mathbf{x}}$. For all $\alpha \in \mathcal{A}$ and conditionally to $\bar{\mathbf{x}}$:

$$\begin{aligned} \|\hat{\mathbf{x}}_\alpha - \bar{\mathbf{x}}\|^2 &= \gamma(\hat{\mathbf{x}}_\alpha, \bar{\mathbf{x}}) - \gamma(\bar{\mathbf{x}}, \bar{\mathbf{x}}) \\ &= [\gamma(\hat{\mathbf{x}}_\alpha, \bar{\mathbf{x}}) - \gamma(\bar{\mathbf{x}}_\alpha, \bar{\mathbf{x}})] + [\gamma(\bar{\mathbf{x}}_\alpha, \bar{\mathbf{x}}) - \gamma(\bar{\mathbf{x}}, \bar{\mathbf{x}})] \\ &= V_\alpha(\bar{\mathbf{x}}) + b_\alpha(\bar{\mathbf{x}}) \end{aligned} \tag{16}$$

where $b_\alpha(\bar{\mathbf{x}}) := \|\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}}\|^2$ is a bias term and $V_\alpha(\bar{\mathbf{x}}) := \gamma(\hat{\mathbf{x}}_\alpha, \bar{\mathbf{x}}) - \gamma(\bar{\mathbf{x}}_\alpha, \bar{\mathbf{x}})$ is a variance term. Note that $V_\alpha(\bar{\mathbf{x}})$ is not equal to $\|\bar{\mathbf{x}}_\alpha - \hat{\mathbf{x}}_\alpha\|^2$ as in the linear case. The bias $b_\alpha(\bar{\mathbf{x}})$ decreases whereas the variance term $V_\alpha(\bar{\mathbf{x}})$ tends to increase when the scale of $\mathcal{C}_\alpha$ increases. Then, taking the expectation of (16) according to $\mathbb{P}_{\bar{\mathbf{x}}}$ leads to

$$\begin{aligned} \mathcal{R}(\bar{\mathbf{x}}, \alpha) &= \mathbb{E}_{\bar{\mathbf{x}}} \|\hat{\mathbf{x}}_\alpha - \bar{\mathbf{x}}\|^2 \\ &= b_\alpha(\bar{\mathbf{x}}) + \mathbb{E}_{\bar{\mathbf{x}}}[V_\alpha(\bar{\mathbf{x}})]. \end{aligned}$$

Using the sampling hypothesis on the $\bar{x}_i$, it yields

$$\mathcal{R}(\mu, \alpha) = \int_{\bar{x}\in\mathcal{G}} b_\alpha(\bar{x})\, d\mu(\bar{x}) + \int_{\bar{x}\in\mathcal{G}} \mathbb{E}_{\bar{x}}[V_\alpha(\bar{x})]\, d\mu(\bar{x}).$$

Conditionally to $\bar{\mathbf{x}}$, the selected model $\hat{\alpha}$ is the one minimizing over the collection the criterion

$$\alpha \mapsto \gamma_n(\hat{\mathbf{x}}_\alpha) + \operatorname{pen}(\alpha). \tag{17}$$

Defining $\hat{b}_\alpha(\bar{\mathbf{x}}) := \gamma_n(\bar{\mathbf{x}}_\alpha) - \gamma_n(\bar{\mathbf{x}})$ and $\widehat{V}_\alpha(\bar{\mathbf{x}}) := \gamma_n(\bar{\mathbf{x}}_\alpha) - \gamma_n(\hat{\mathbf{x}}_\alpha)$, the selected model is a minimizer of

$$\begin{aligned} \gamma_n(\hat{\mathbf{x}}_\alpha) - \gamma_n(\bar{\mathbf{x}}) + \operatorname{pen}(\alpha) &= \gamma_n(\hat{\mathbf{x}}_\alpha) - \gamma_n(\bar{\mathbf{x}}_\alpha) + \gamma_n(\bar{\mathbf{x}}_\alpha) - \gamma_n(\bar{\mathbf{x}}) + \operatorname{pen}(\alpha) \\ &= \hat{b}_\alpha(\bar{\mathbf{x}}) - \widehat{V}_\alpha(\bar{\mathbf{x}}) + \operatorname{pen}(\alpha). \end{aligned} \tag{18}$$

Then, introducing the term of interest (16) into (18), it leads to

$$\begin{aligned} \gamma_n(\hat{\mathbf{x}}_\alpha) - \gamma_n(\bar{\mathbf{x}}) + \operatorname{pen}(\alpha) &= b_\alpha(\bar{\mathbf{x}}) + V_\alpha(\bar{\mathbf{x}}) + \left[\hat{b}_\alpha(\bar{\mathbf{x}}) - b_\alpha(\bar{\mathbf{x}})\right] - \left[V_\alpha(\bar{\mathbf{x}}) + \widehat{V}_\alpha(\bar{\mathbf{x}})\right] + \operatorname{pen}(\alpha) \\ &= \|\hat{\mathbf{x}}_\alpha - \bar{\mathbf{x}}\|^2 + \left[\hat{b}_\alpha(\bar{\mathbf{x}}) - b_\alpha(\bar{\mathbf{x}})\right] - \left[V_\alpha(\bar{\mathbf{x}}) + \widehat{V}_\alpha(\bar{\mathbf{x}})\right] + \operatorname{pen}(\alpha). \end{aligned}$$

Some concentration arguments allows us to suppose that $\|\hat{\mathbf{x}}_\alpha - \bar{\mathbf{x}}\|^2$ is close to $\mathcal{R}(\bar{\mathbf{x}}, \alpha)$ (see [28] p. 9) for n large enough. Furthermore, the law of large numbers for the distribution of $\bar{\mathbf{x}}$ yields that

$$\mathcal{R}(\bar{\mathbf{x}}, \alpha) \to \mathcal{R}(\mu, \alpha) = \int_{\bar{x}\in\mathcal{G}} \mathcal{R}(\bar{\mathbf{x}}, \alpha)\, d\mu(\bar{x})$$

when n tends to infinity. Thus $\|\hat{\mathbf{x}}_\alpha - \bar{\mathbf{x}}\|^2 \approx \mathcal{R}(\mu, \alpha)$ if n is large enough. Next, we can easily show that

$$b_\alpha(\bar{\mathbf{x}}) - \hat{b}_\alpha(\bar{\mathbf{x}}) = 2\sigma\langle \bar{\mathbf{x}} - \bar{\mathbf{x}}_\alpha, \xi\rangle$$

where ξ is a random vector independent of the random vectors $\bar{\mathbf{x}}$ and $\bar{\mathbf{x}}_\alpha$. The law of large numbers shows that $\langle \bar{\mathbf{x}} - \bar{\mathbf{x}}_\alpha, \xi\rangle$ tends to 0 when n tends to infinity. Thus,

$$\gamma_n(\hat{\mathbf{x}}_\alpha) - \gamma_n(\bar{\mathbf{x}}) + \mathrm{pen}(\alpha) \approx \mathcal{R}(\mu, \alpha) - \left[V_\alpha(\bar{\mathbf{x}}) + \widehat{V}_\alpha(\bar{\mathbf{x}})\right] + \mathrm{pen}(\alpha). \tag{19}$$

In order to make (19) close to the integrated risk $\mathcal{R}(\mu, \alpha)$, the *optimal penalty* is defined by

$$\mathrm{pen}_{\mathrm{opt}}(\alpha) = V_\alpha(\bar{\mathbf{x}}) + \widehat{V}_\alpha(\bar{\mathbf{x}}).$$

Next, the main point of this heuristics is to assume that $\widehat{V}_\alpha(\bar{\mathbf{x}}) \approx V_\alpha(\bar{\mathbf{x}})$. Finally, this assumption leads to $\mathrm{pen}_{\mathrm{opt}}(\alpha) = 2\widehat{V}_\alpha(\bar{\mathbf{x}})$. Turning back on the expression of $\widehat{V}_\alpha(\bar{\mathbf{x}})$, it can be written

$$\begin{aligned} \widehat{V}_\alpha(\bar{\mathbf{x}}) &= \gamma_n(\bar{\mathbf{x}}_\alpha) - \gamma_n(\bar{\mathbf{x}}) + \gamma_n(\bar{\mathbf{x}}) - \gamma_n(\hat{\mathbf{x}}_\alpha) \\ &= \hat{b}_\alpha(\bar{\mathbf{x}}) + \gamma_n(\bar{\mathbf{x}}) - \gamma_n(\hat{\mathbf{x}}_\alpha). \end{aligned}$$

For large α, the bias term stabilizes itself since the approximation of the model cannot be appreciably improved. Thus, the behaviour of $\widehat{V}_\alpha(\bar{\mathbf{x}})$ according to $\ln|\mathcal{C}_\alpha|_k$ is known for large α via $-\gamma_n(\hat{\mathbf{x}}_\alpha)$. In our framework, with a fixed observation sample, the penalty functions could be regarded as proportional to $\ln|\mathcal{C}_\alpha|_k$. Next,

$$\mathrm{pen}_{\mathrm{opt}}(\alpha) = 2\widehat{V}_\alpha(\bar{\mathbf{x}}) = -2\beta_{\mathrm{opt}}\ \ln|\mathcal{C}_\alpha|_k$$

where β_{opt} is a constant. In order to use the slope heuristics to calibrate the penalty, a required condition is to observe a linear trend in the point cloud $\{\ln|\mathcal{C}_\alpha|_k, SS(\alpha)\}_{\alpha\in\mathcal{A}}$ for large α where $SS(\alpha) := \|\hat{\mathbf{x}}_\alpha - \mathbf{X}\|^2 = \gamma_n(\hat{\mathbf{x}}_\alpha) + \|\mathbf{X}\|^2$. If this condition is fulfilled, an estimator $\hat{\beta}$ of β_{opt} can be computed by regressing $SS(\alpha)$ on $\ln|\mathcal{C}_\alpha|_k$ for large α and the final penalty is

$$\mathrm{pen}(\alpha) = -2\hat{\beta}\ln|\mathcal{C}_\alpha|_k.$$

5 Experimental results

The experimental studies presented in this section deal with 1-skeleton of α-shape complexes, namely α-shape graphs. They aim at illustrating our method in some simple situations. A discussion about the use of other complexes is also given in Section 6.

Our objective is to study a set of points using simplicial complexes. We first define a set of landmark points form the observed points, and then a collection of simplicial complexes is defined on these landmarks points. Finally, the slope heuristics is used to select a simplicial complex in the collection.

5.1 Ideal complexes

For the applications, the choice of a family of "good" landmarks is of first importance to make our procedure operational. We start with some ideal models in the sense that the complexes used are built on some landmarks which are really located on a geometric object. The aim of this first application subsection is twofold : first this one allows us to detail the complex selection procedure in practice, and second it shows that a "slope behaviour" of the family of least squares estimators can be observed in this simple framework.

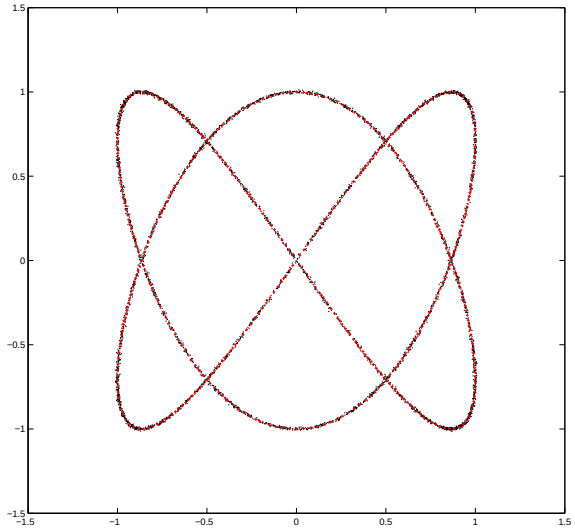
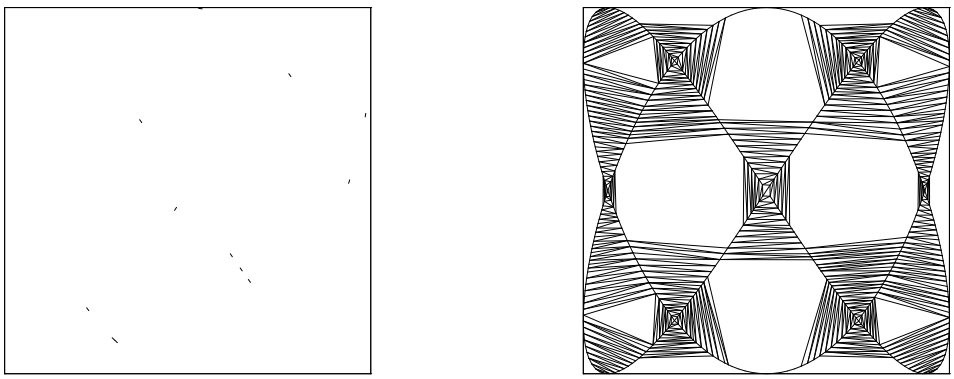


Figure 1: Lissajous curve simulation. Red points are landmarks and black one are observed points.



(a) Thinner complex of the collection : $\alpha = 0.00015$.

(b) Thicker complex of the collection : $\alpha = 0.1$.

Figure 2: The two extremal complexes of the considered complex family. The bias of the thinner complex (a) is too large whereas the thicker complex (b) suffers from too much overfitting.

Lissajous curve simulation

We first study a Lissajous curve $\mathcal{G}$ parametred in $\mathbb{R}^2$ according to

$$t \in [0, \pi], \quad \begin{cases} \bar{x}^{(1)}(t) = \sin 2t \\ \bar{x}^{(2)}(t) = \sin 3t \end{cases} .$$

1-Complex family :

- *Landmarks generation :* A set $\mathcal{P}_l$ of 10000 points is (exactly) sampled on the Lissajous curve by sampling uniformly some points t in $[0, \pi]$, this set of point is dedicated to the determination of a subset of landmarks. A *furthest point* rule is used to extract the subset of the landmark on this family. More precisely, we choose a first landmark z_1 at random in $\mathcal{P}_l$. The second landmark z_2 is then the furthest point of z_1 in $\mathcal{P}_l$, z_3 is the furthest point of $\{z_1, z_2\}$ in $\mathcal{P}_l$, etc... until we have defined $p = 500$ landmarks.
- *Complexes :* We use some α-complexes $\mathcal{C}_\alpha$ of dimension 1 (graphs) built on the landmark family $\{z_1, \ldots, z_p\}$; these complexes are computed using the C++ CGAL library[8]. The family of critical parameters α for which the complexes change is automatically provided by CGAL. The largest α (greater than 0.1 for instance in this example) are removed since they correspond to huge complexes, see Figure 2b for an illustration. The set $\mathcal{A}$ of remaining coefficients α defines the considered family of graphs. The length $l(\alpha)$ of each graph $\mathcal{C}_\alpha$ is also computed.

2-Observation points :

- Another set of "true points" $\bar{x}_1, \ldots, \bar{x}_n$ is generated in the same way on the Lissajous curve. The sample of observed points $\mathcal{P}_o = \{X_1, \ldots, X_n\}$ is

$$\forall i = 1, \ldots, n, \quad X_i = \bar{x}_i + \sigma \xi_i$$

  with $n = 5000$ and where the ξ_i are iid standard Gaussian vectors of $\mathbb{R}^2$ and $\sigma = 0.005$ is the noise level.

3-Estimations :

- *Sum of squares :* For each complex $\mathcal{C}_\alpha$ in the family, the estimator $\hat{x}_{\alpha,i}$ is the closest point of X_i which belongs to $\mathcal{C}_\alpha$. Since $\mathcal{C}_\alpha$ is a graph, this computation only involves projection on segments. The squared distances $SS(\alpha)$ between the observation vector $\mathbf{X}$ and the estimator vector $\hat{\mathbf{x}}_\alpha$ is also computed.

4-Slope heuristics - complex selection :

- *Validation :* Check that the point could $\{\ln l(\alpha), SS(\alpha)\}_{\alpha \in \mathcal{M}}$ has a linear trend for α large enough.
- *Penalty definition :* Compute the slope coefficient $\hat{\beta}$ for the regression of $SS(\alpha)$ according to $\ln l(\alpha)$. The final penalty is $\mathrm{pen}(\alpha) = -2\hat{\beta} \ln l(\alpha)$.
- *Complex selection :* The selected complex is $\mathcal{C}_{\hat{\alpha}}$ where

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha \in \mathcal{M}} SS(\alpha) + \mathrm{pen}(\alpha).$$

Note that the framework detailed in these four steps exactly corresponds to the hypotheses of Theorem 2. Thus, we are in a perfect position to procede the complex selection method. For this simulated dataset, the true positions $\bar{\mathbf{x}}$ on the Lissajous curve are known, thus the risk can be estimated by a Monte-Carlo procedure : the steps 2, 3 and 4 are repeated for a fixed complex family, the distance $\|\bar{\mathbf{x}} - \hat{\mathbf{x}}_\alpha\|^2$ is computed each time and finally the mean of these quantities

leads to an approximation of the risk $\mathcal{R}(\bar{\mathbf{x}}, \alpha)$ of the estimator $\hat{\mathbf{x}}_\alpha$. This procedure allows us to evaluate the performance of the model selection method.

Figure 3 summarizes the results of the Lissajous simulation study. The oracle complex printed on Figure 3a corresponds to the minimum risk plotted in Figure 3c, more precisely $\alpha_{\text{or}} = 0.001256$ and $\ln l(\alpha_{\text{or}}) = 2.855$. In comparison with the two extremal complexes of Figure 2, the oracle is close to the true Lissajous curve and thus it makes sense to track it with a model selection procedure. As shown in Figure 3c, the slope behaviour is actually observed for large α. The estimation of this slope for $\alpha \in [0.00255\, 0.0652]$ leads to the selected complex 3b with $\hat{\alpha} = 0.001286$ and $\ln l(\hat{\alpha}) = 2.862$. Thus, the oracle complex and the selected one are close to each other. Indeed, no obvious differences can be seen between the figures 3a and 3b.

$\alpha \quad \times 10^{-3}$	$\alpha_{\min}$...	1.129	1.255	1.256	1.283	1.286	1.298	1.344
$N(\alpha)$	0	1	369	6	19	77	3	10
Selection percentage	0	0.2	73.8	1.2	3.8	15.4	0.6	2
Length	0.0394	16.86	17.30	17.37	17.45	17.50	17.57	17.64
Risk $\times 10^{-5}$	29841	2.627	2.589	2.588	2.591	2.594	2.594	2.596

$\alpha \quad \times 10^{-3}$	1.493	1.603	1.643	1.669	1.672	1.748	...$\alpha_{\max}$
$N(\alpha)$	6	4	1	2	1	1	0
Selection perc.	1.2	0.8	0.2	0.4	0.2	0.2	0
Length	17.71	17.97	18.10	18.31	18.46	18.61	185.8
Risk $\times 10^{-5}$	2.606	2.618	2.623	2.639	2.641	2.642	3.946

Table 1: Number of times $N(\alpha)$ each graph $\mathcal{C}_\alpha$ in the collection is selected over 500 identical experiences associated to the Lissajous curve. The red column corresponds to the oracle complex.

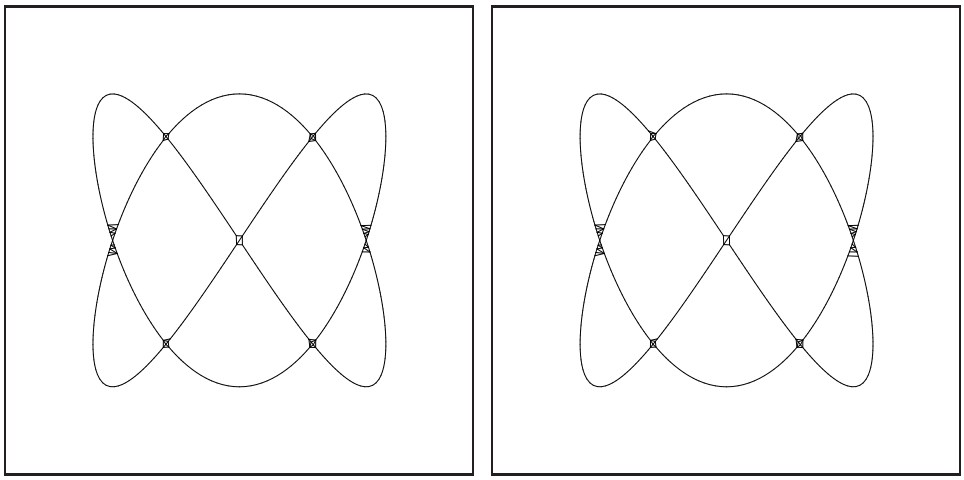
In order to evaluate the behaviour of the procedure, the sample $\mathcal{P}_o = \{X_1, \ldots, X_n\}$ is simulated 500 times. The true locations $\bar{x}_i$ are unchanged as well as the complex collection. The regression for the slope estimation is proceeded for $\alpha \in [0.00255\,, 0.0652]$ and for then we count the number of times each model is selected over the 500 experiences. Results are given in Table 1, it shows that the complex just before ($\alpha = 0.001255$) is selected about 3 times over 4. This complex has just one edge less than the oracle and its risk is nearly the same as the oracle one. Thus we can conclude that for this simulation, the penalized criterion succeeds in selecting good complexes with a risk close to the oracle one.

Spiral and segment

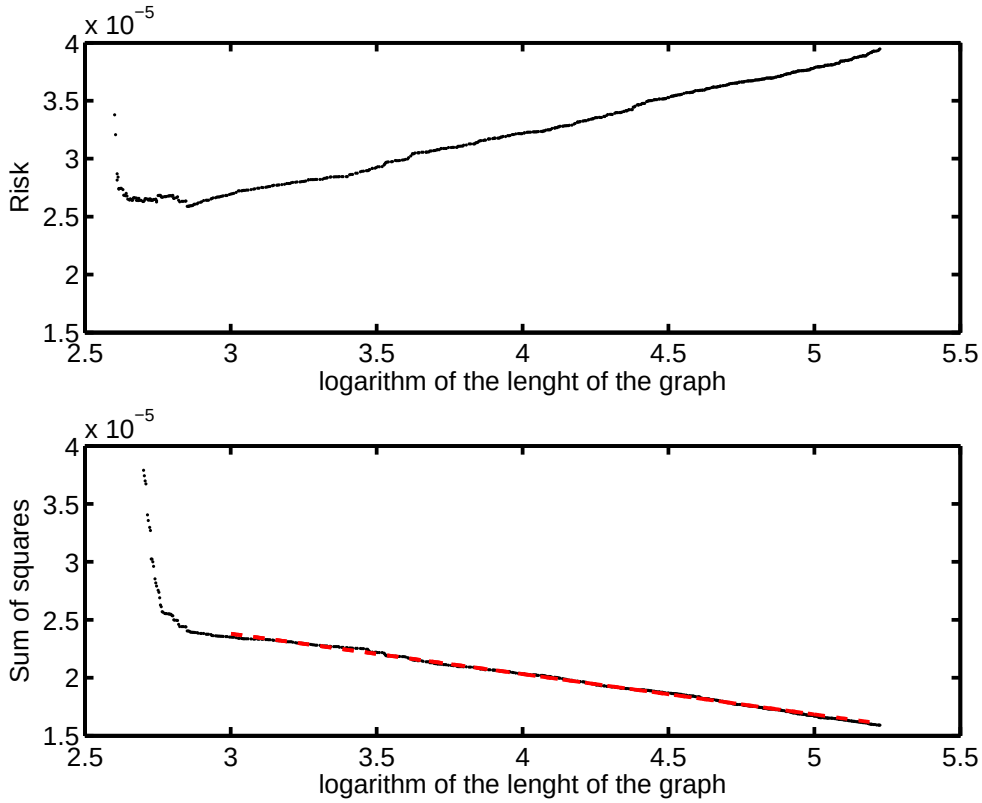
This second example is the union of a spiral and a segment in $\mathbb{R}^2$. As in the last simulation, $p = 120$ landmarks are extracted from an initial set of points $\mathcal{P}_l$ sampled on these two elements with no additional noise, whereas it is the case for the "observed points" with a level noise $\sigma = 0.01$. The sampling is uniform along the spiral and the segment. Figure 4(a) shows the landmarks and a sample of $n = 1000$ observed points. The steps 1 to 4 detailed in the previous section are followed in order to define a complex family and to select one according to the slope heuristics.

The results of the slope heuristics procedure are summarized on Figure 4. Once again, the slope behaviour is observed as shown on the bottom plot of Figure 4c. For this particular simulation, the selected complex is exactly the oracle complex plotted on Figure 4b, with $\alpha_{\text{or}} = \hat{\alpha} = 0.0006121$ and $\ln l(\alpha_{\text{or}}) = \ln l(\hat{\alpha}) = 1.247$. As expected, the spiral and the segment are easily separated.

As in the last section, the behaviour of the slope procedure is studied on 500 simulations of $\mathcal{P}_o = \{X_1, \ldots, X_n\}$. As before the fixed true locations $\bar{\mathbf{x}}_i$ and the complex collection are unchanged for all the experiences. The regressions have been proceeded with $\alpha \in [0.0124\ \ 0.0931]$ for each experience. Table 2 shows that the selected complex is most of the time close to the oracle one. The two following coefficients α after α_{or} are more than ten times greater then α_{or}, but the two next graphs after $\mathcal{C}_{\text{or}}$ are not much larger than $\mathcal{C}_{\text{or}}$ and their risks are really close to the oracle. The results are satisfying since the selected complex is one of these three graphs in nearly 70% of the cases.

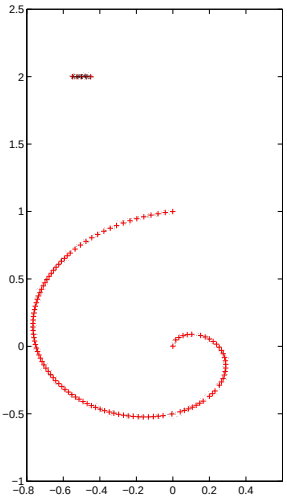


(a) Oracle complex. (b) Selected complex.

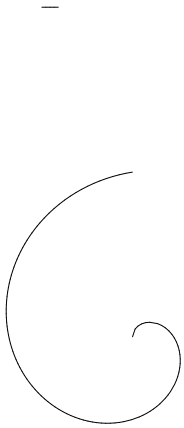


(c) Risk and sum of squares $\mathrm{SS}(\alpha)$. These curves are zooms around the minimum of the risk.

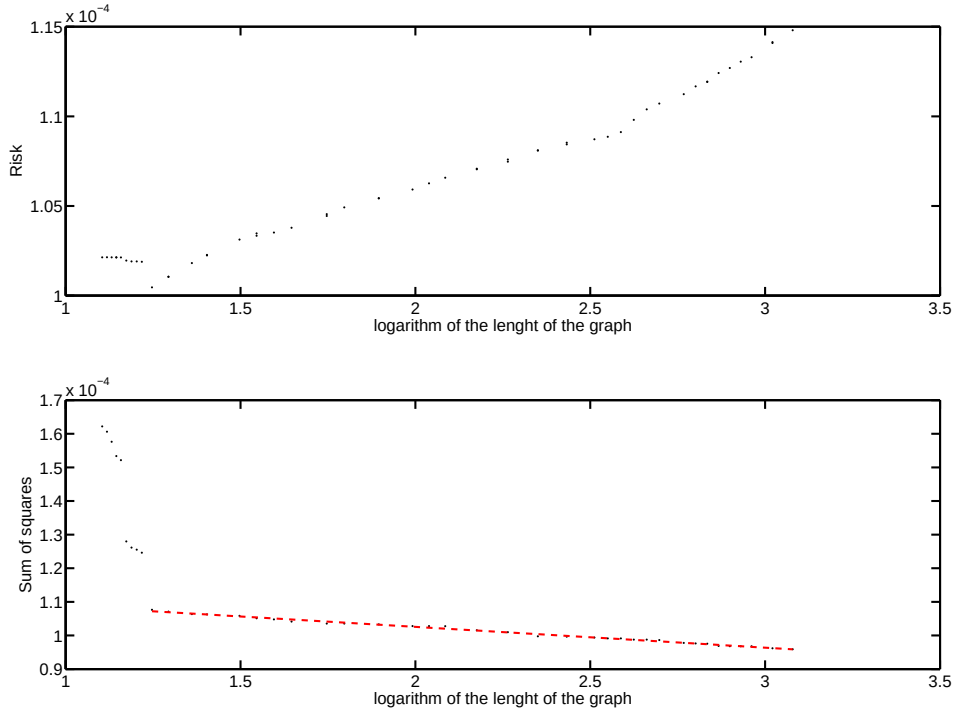
Figure 3: Results of the slope heuristics for the selection of an "ideal" complex for the Lissajous curve simulation. The regression for the slope estimation has been done on the interval corresponding to $\alpha \in [0.00255, 0.0652]$. We have $\alpha_{\text{or}} = 0.001256$, $\ln l(\alpha_{\text{or}}) = 2.855$, $\hat{\alpha} = 0.001286$ and for this simulation $\ln l(\hat{\alpha}) = 2.862$.



(a) Landmarks and observed points.



(b) Oracle and selected complex.



(c) Risk and sum of squares $\mathrm{SS}(\alpha)$. These curves are zooms around the minimum of the risk.

Figure 4: Illustration of the slope heuristics for the selection of an "ideal" complex for the Spiral-segment example with $n = 1000$ observation points and $p = 120$ landmarks. The regression for the slope estimation has been done on the interval corresponding to $\alpha \in [0.01243\ 0.09317]$. For this particular simulation, the selected complex is exactly the oracle complex , with $\alpha_{\mathrm{or}} = \hat{\alpha} = 0.0006121$ and $\ln l(\alpha_{\mathrm{or}}) = \ln l(\hat{\alpha}) = 1.247$.

$\alpha \quad \times 10^{-2}$	$\alpha_{\min}$...	0.0612	0.8186	1.042	1.243	1.63
$N(\alpha)$	0	64	152	127	57	52
Selection percentage	0	12.8	30.4	25.4	11.4	10.4
Length	0.02458	3.479	3.648	3.900	4.071	4.467
Risk $\times 10^{-4}$	20457	1.005	1.011	1.018	1.023	1.031

$\alpha \quad \times 10^{-2}$	1.809	2.168	2.351	2.537	2.913	3.100	...$\alpha_{\max}$
$N(\alpha)$	28	11	1	6	1	1	0
Selection perc.	5.6	2.2	0.2	1.2	0.2	0.2	0
Length	4.692	4.931	5.185	5.735	6.030	6.656	21.73
Risk $\times 10^{-4}$	1.035	1.0352	1.038	1.044	1.049	1.054	1.148

Table 2: Number of times $N(\alpha)$ each graph $\mathcal{C}_\alpha$ in the collection is selected over 500 identical experiences associated to the spiral-segment simulation. The red column corresponds to the oracle complex.

$\alpha \quad \times 10^{-3}$	$\alpha_{\min}$...	0.9537	0.9891	1.051	1.076	1.078	1.084
$N(\alpha)$	0	38	3	107	36	281	2
Selection percentage	0	7.6	0.6	21.4	7.2	56.2	0.4
Length	0.03083	17.45	17.64	17.87	17.97	18.02	18.09
Risk $\times 10^{-4}$	308	1.1910	1.1899	1.1897	1.1942	1.1939	1.1937

$\alpha \quad \times 10^{-3}$	1.126	1.183	1.187	1.200	1.205	1.271	...$\alpha_{\max}$
$N(\alpha)$	13	12	0	4	1	3	0
Selection perc.	2.6	2.4	0	0.8	0.2	0.6	0
Length	18.29	18.34	18.38	18.49	18.55	18.82	146.1
Risk $\times 10^{-4}$	1.1898	1.1886	1.1885	1.1899	1.1932	1.1944	1.6823

Table 3: Number of times $N(\alpha)$ each graph $\mathcal{C}_\alpha$ in the collection is selected over 500 identical experiences associated to the Lissajous curve simulation, with "noised" landmarks. The red column corresponds to the oracle complex.

5.2 Landmarks selection from observed dataset

Landmark choice

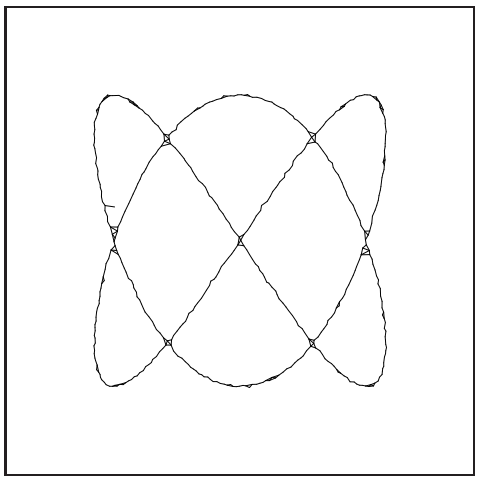
Roughly speaking, if the landmarks are far from $\mathcal{A}$, the approximation ability of the complexes built on these landmark points will be affected. Several methods can be tested to define convenient landmarks, namely landmarks which allow us to apply the slope heuristics. Choosing landarks randomly in $\mathcal{P}_l$ leaves some large area without any landmarks if they are not numerous enough. The *furthest point* strategy tends to focus on extremal points or even outliers, which is exactly the opposite of what we need. The standard k-means clustering algorithm[25] fills better to our purpose since it aims at finding some prototypes $z_1, \ldots, z_p$ which minimize

$$(z_1, \ldots, z_p) \longmapsto \sum_{x \in \mathcal{P}_l} \min_{i=1,\ldots,p} \|x - z_i\|^2.$$

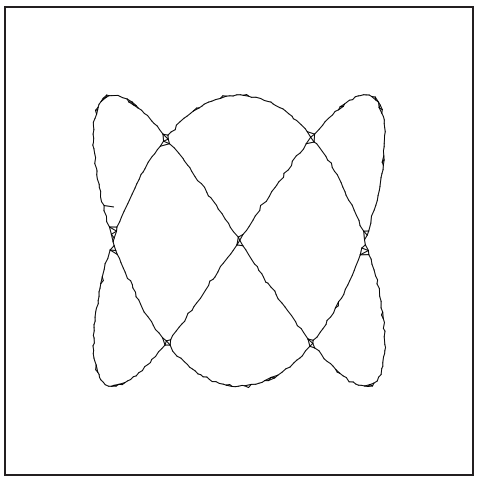
The "Neural-Gas" algorithm proposed by [27] is an extension of the k-means algorithm to avoid confinement to local minima. In the following, the landmarks are defined thanks to this method which has already been used in [3] with a similar framework.

Lissajous curve

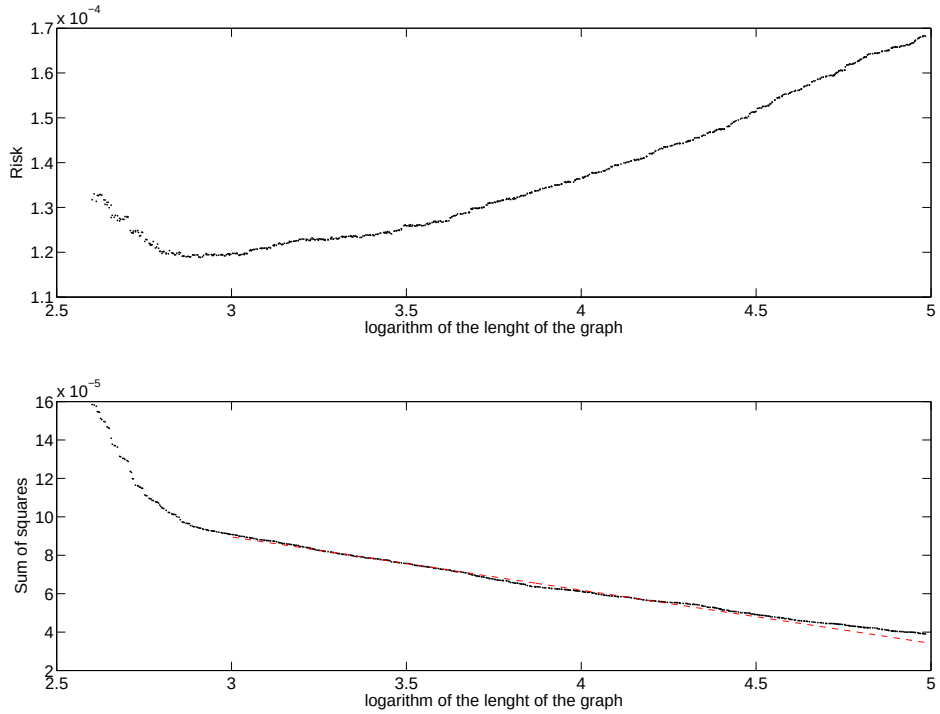
We return to the Lissajous curve example. The complete point set $\mathcal{P}$ is randomly separated into $\mathcal{P}_o$ (5000 points) and $\mathcal{P}_l$ (5000 points). As before, the set $\mathcal{P}_l$ is used to produce a set of



(a) Oracle complex.



(b) Selected complex.



(c) Risk and sum of squares $\mathrm{SS}(\alpha)$. These curves are zooms around the minimum of the risk.

Figure 5: Results of the slope heuristics for the selection of a complex defined on noisy landmarks for the Lissajous curve simulation. The regression for the slope estimation has been done on the interval corresponding to $\alpha \in [0.00607\ 0.034]$. We found $\alpha_{\mathrm{or}} = 0.001187$, $\ln l(\alpha_{\mathrm{or}}) = 2.911$, and for this particular simulation $\hat{\alpha} = 0.001051$ and $\ln l(\hat{\alpha}) = 2.883$.

$p = 500$ landmarks denoted $z_1, \ldots, z_p$. The landmarks are obtained thanks to the neural gas algorithm and the α-complexes are built on these landmarks. Next, the steps 2 to 4 are followed as before. The regression for the slope estimation has been done on the interval corresponding to $\alpha \in [0.00607\ 0.034]$. We found $\alpha_{\text{or}} = 0.001187$ and $\ln l(\alpha_{\text{or}}) = 2.911$. For the particular simulation presented on Figure 5, we found $\hat{\alpha} = 0.001051$ and $\ln l(\hat{\alpha}) = 2.883$. Note that the length of the oracle complex is larger than when the position of the landmarks are truly on the Lissajous curve. Indeed, the landmarks are probably not as efficient as if they were on the curve and thus the oracle graph has to be enlarged. Table 3 shows the results for 500 repetitions of the same simulation and the same application of the slope method. Although the results are not as good as in the previous section, the selected simplicial complexes still have correct risk performances.

5.3 Seismic data

In this section, we show that the slope heuristics method behaves well even with a real dataset. The Centennial Catalog [19] is a global catalog of instrumentally recorded earthquakes. This catalog provides the locations and magnitudes of large earthquakes since 1900, the data can be downloaded from the USGS website[1]. Here, we are only interested in the location of the earthquakes and we do not study the magnitude. Figure 6a shows a picture of the seismicity distribution in the Earth. We intent to infer the geological faults from the earthquake data with α-shape graphs.

As in the previous simulations, a landmark family of 1000 points is determined thanks to the neural gas algorithm. In addition, the more isolated points have been removed from the landmark family. Indeed, if the isolated landmarks are kept, the isolated landmarks are actually vertices of the graphs only for largest α. In this case, this fact implies that LSE criterion shows a sequence of substantial jumps (the bias still decreases), even for the largest complexes of the collection and the slope heuristics cannot be applied. In the particular example presented in Figure 6, the 5% points the most isolated have been removed from the original landmark family, namely the 5% landmark points with the largest distance to their first neighbour. At the end, the landmark family is composed of 950 points.

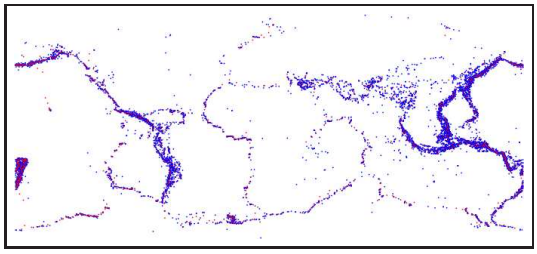
Of course, the oracle graph is unknown in this real example, but Figure 6 shows that the LSE criterion has a slope behavior and thus the slope method can be used to select a graph in the collection. The selected graph provides a representation of the geological faults on Earth. Note that the interest of this example is more in showing that a slope behavior can be observed on real data than in giving an efficient model for the seismic activity.

6 Discussion

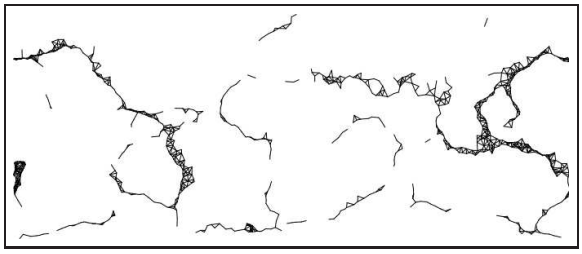
This work should be considered as a first attempt to use elaborated statistical methods for geometrical purposes. We propose to introduce model selection tools in the computational geometry field to choose a simplicial complex in a given collection of homogeneous simplicial complexes. The minimization of a penalized least squares criterion leads to the selection of a simplicial complex. In order to minimize the risk of the estimators associated to the complex collection, it is shown that the penalty has to be chosen proportional to $|\mathcal{C}_\alpha|$. Next, the slope heuristics method is used in practice to calibrate the penalty from the data. In this paper, the model selection procedure is precisely studied in the particular case of α-shape graphs. Applications to more elaborate scenarios with complexes of higher dimensions will be studied in forthcoming work.

Note that the experimentation proposed in Section 5 does not exactly correspond to the hypotheses of Theorem 2 since the computed complexes necessary depend on the observed data and thus are "not fixed" as in the theorem statement. Our theoretical result can be considered as conditionally to the landmark choice. Giving some mathematical results for the "random models" we use would be much more difficult among other things because the distribution of the landmarks cannot be easily specified... Nevertheless, the neural-gas algorithm tends to converge to

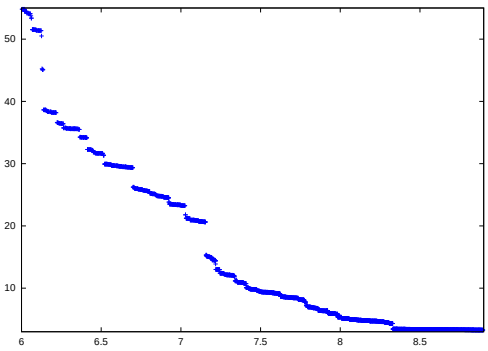
[1]The dataset can be downloaded at `http://earthquake.usgs.gov/research/data/centennial.php`



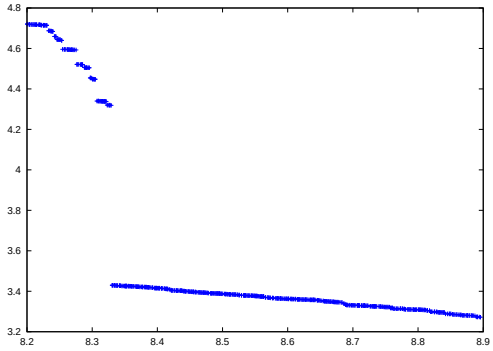
(a) Landmarks(red) and observed points(blue).



(b) Selected graph.



(c) LSE criterion versus logarithm of the graph length.



(d) LSE criterion versus logarithm of the graph length (zoom for large α).

Figure 6: Results of the slope heuristics for the selection of a complex for the seismic data.

some "optimal points" that depends on the distribution of the points in $\mathcal{P}_l$ and thus the situation we deal with is not so far from the theoretical results given in Section 3.2.

In practice, all the simplicial complexes does not allow us to use the slope heuristics. For instance with Rips complexes, no linear behaviour of the least square criterion can be observed in general. One possible explanation for this phenomenon is that the Rips complexes are too rich, namely they create too many simplices. With an approximation point of view, Rips graphs need much more segments than α-shape graphs to provide a good approximation of a geometric object. Actually, little is known about the approximation properties of these different complex families. An approximation theory of simplicial complexes would be really helpful to describe how the bias decreases as the complex scale increases.

Topological properties of an unknown object can be inferred from a point cloud sampled on it by several methods. One possible strategy is to consider the topology of an union of balls centered on theses points [see for instance 12, 11]. Roughly speaking, our complex selection method provides a "convenient scale" at which the geometric features has to be studied. Thus, the selected $\hat{\alpha}$ can be taken for the radius of the balls in order to recover the topology of the underlying object.

The authors are grateful to Frédéric Chazal (INRIA Saclay), Steve Oudot (INRIA Saclay), Pascal Massart (Université Paris-Sud 11) and Primoz Skraba (INRIA Saclay) for helpful discussions and valuable comments on both mathematical and computational aspects of this work.

A Proof of Theorem 2

Proof. Let $\mathcal{C}_\alpha$ be a simplicial model of the collection. According to Proposition 1, for all $r \leq \delta_{C_\alpha}$,

$$\begin{aligned}
\Phi_\alpha(r) = \int_0^r \sqrt{H(\mathcal{C}_\alpha^n, \|\cdot\|, t)}dt &\leq \int_0^r \sqrt{nk \ln \frac{4|\mathcal{C}_\alpha|_k}{t}}dt \\
&\leq \sqrt{nk}\, 4|\mathcal{C}_\alpha|_k \int_0^{r/4|\mathcal{C}_\alpha|_k} \sqrt{\ln \frac{1}{u}}du \\
&\leq \sqrt{nk}\, 4|\mathcal{C}_\alpha|_k \frac{r}{4|\mathcal{C}_\alpha|_k} \left\{ \sqrt{\ln \frac{4|\mathcal{C}_\alpha|_k}{r}} + \sqrt{\pi} \right\} \\
&\leq r\sqrt{nk} \left\{ \sqrt{\ln \frac{4|\mathcal{C}_\alpha|_k}{r}} + \sqrt{\pi} \right\}.
\end{aligned}$$

Indeed, for all $r \in (0, 1]$,

$$\int_0^r \sqrt{\ln\left(\frac{1}{x}\right)}\, dx \leq r \left\{ \sqrt{\ln\left(\frac{1}{r}\right)} + \sqrt{\pi} \right\}.$$

Thus, for all $r \leq \delta_{C_\alpha}$, we set

$$\varphi_\alpha(r) := \kappa r\sqrt{nk} \left\{ \sqrt{\ln \frac{4|\mathcal{C}_\alpha|_k}{r}} + \sqrt{\pi} \right\}$$

where κ is the same constant as in Section 3.1. If $r \geq \delta_{C_\alpha}$ then we can set $\varphi_\alpha(r) = \varphi_\alpha(\delta_{C_\alpha}) + \varphi_\alpha(\delta_{C_\alpha})(r - \delta_{C_\alpha})$. Then it is clear that $\Phi_\alpha(r) \leq \varphi_\alpha(r)$.

The next set to apply Theorem 1 consists in finding a solution to the equation

$$\varphi_\alpha\left(2\frac{\sigma}{\sqrt{nD}}\sqrt{d_\alpha}\right) = \frac{\sigma}{\sqrt{nD}} d_\alpha. \tag{20}$$

This is equivalent to solve $\Phi_\alpha(r) = \frac{\sqrt{nD}}{4\sigma} r^2$ and there is a unique solution $r_\alpha = 2\sigma\sqrt{\frac{d_\alpha}{nD}}$ to this last equation because of the concavity of φ_α and the convexity of $r \mapsto r^2$. Next,

$$\begin{aligned}
r_\alpha \leq \delta_{\mathcal{C}_\alpha} &\Leftrightarrow \kappa\delta_{\mathcal{C}_\alpha}\sqrt{nk} \left\{ \sqrt{\ln \frac{4|\mathcal{C}_\alpha|_k}{\delta_{\mathcal{C}_\alpha}}} + \sqrt{\pi} \right\} < \frac{\sqrt{nD}}{4\sigma} \delta_{\mathcal{C}_\alpha}^2 \\
&\Leftrightarrow \sigma \leq \delta_{\mathcal{C}_\alpha} \sqrt{\frac{D}{k}} \left[4\kappa \left(\sqrt{\ln \frac{4|\mathcal{C}_\alpha|_k}{\delta_{\mathcal{C}_\alpha}}} + \sqrt{\pi} \right) \right]^{-1}
\end{aligned} \tag{21}$$

Under Condition (21), then

$$\begin{aligned}
(20) &\Leftrightarrow \kappa r_\alpha \sqrt{nk} \left\{ \sqrt{\ln \frac{4|\mathcal{C}_\alpha|_k}{r_\alpha}} + \sqrt{\pi} \right\} = \frac{\sqrt{nD}}{4\sigma} r_\alpha^2 \\
&\Leftrightarrow 4\kappa\sigma\sqrt{\frac{k}{D}} \left\{ \sqrt{\ln \frac{4|\mathcal{C}_\alpha|_k}{r_\alpha}} + \sqrt{\pi} \right\} = r_\alpha.
\end{aligned}$$

Thus, $4\kappa\sigma\sqrt{k\pi}/\sqrt{D} \leq r_\alpha$ and then

$$4\kappa\sigma\sqrt{\frac{k}{D}} \left\{ \sqrt{\ln \frac{|\mathcal{C}_\alpha|_k\sqrt{D}}{\kappa\sigma\sqrt{k\pi}}} + \sqrt{\pi} \right\} \geq r_\alpha$$

or equivalently

$$d_\alpha \leq 8\kappa^2 nk \left\{ \ln \frac{|\mathcal{C}_\alpha|_k \sqrt{D}}{\kappa\sigma\sqrt{k\pi}} + \pi \right\}.$$

□

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[2] S. Arlot and P Massart. Data-driven calibration of penalties for least-squares regression. *J.Mach.Learn.Res.*, 10:245–279, 2009.

[3] Michaël Aupetit. Learning topology with the generative gaussian graph and the em algorithm. In *Advances in Neural Information Processing Systems*, 2006.

[4] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

[5] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3:203–268, 2001.

[6] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138:33–73, 2007.

[7] Christopher M. Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York, 2006.

[8] CGAL Editorial Board. *CGAL User and Reference Manual*, 3.4 edition, 2008.

[9] J.D. Boissonnat, L.J. Guibas, and S. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. *Proc. 23rd ACM Sympos. on Comput. Geom.*, 197:194–203, 2007.

[10] K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York, 2nd edition, 2002.

[11] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in euclidean spaces. *Discrete Comput Geom*, 41:461, 2009.

[12] F. Chazal and A. Lieutier. Smooth manifold reconstruction from noisy and non uniform approximation with guarantees. *Comp. Geom: Theory and Applications*, 40:156, 2008.

[13] F. Chazal and S. Oudot. Towards persistence-based reconstruction in euclidean spaces. In *Proc. 24th ACM Sympos. on Comput. Geom.*, pages 232–241, 2008.

[14] Siu-Wing Cheng and Man-Kwun Chiu. Dimension detection via slivers. In *SODA 09: ACM -SIAM Symposium on Discrete Algorithms*, pages 1001–1010, 2009.

[15] Vin de Silva. A weak characterisation of the delaunay triangulation. *Geometriae Dedicata*, 135, 2008.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B.*, 39:1–38, 1977.

[17] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.

[18] H. Edelsbrunner and E.P. Mücke. Three-dimensional alpha shapes. *ACM Transactions on Graphics*, 13:43–72, 1994.

[19] E.R. Engdahl and A. Villaseñor. *Global Seismicity: 1900–1999, Part A International Handbook of Earthquake and Engineering Seismology*, chapter 41, page pp. 665–690. Academic Press, 2002.

[20] P. Gaillard, M. Aupetit, and G. Govaert. Learning topology of a labeled data set with the supervised generative gaussian graph. *Neurocomputing*, 71, 2008.

[21] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[22] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[23] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2001.

[24] E. Lebarbier. Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85:717–736, 2005.

[25] J. B. Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[26] Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

[27] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten. 'neural-gas' network for vector quantization and its application to time-series prediction. *Neural Networks, IEEE Transactions on*, 4:558–569, 1993.

[28] Pascal Massart. *Concentration Inequalities and Model Selection*, volume Lecture Notes in Mathematics. Springer-Verlag, 2007.

[29] C. Maugis and B. Michel. Slope heuristics for variable selection and clustering via Gaussian mixtures. Technical Report 6550, INRIA, 2008.

[30] P. Niyogi, S. Smale, and S. Weinberger. A topological view of unsupervised learning and clustering. Technical Report TR-2008-01, Computer Science Dept., University of Chicago.

[31] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, March 2008.

[32] Gilles Pisier. *The volume of convex bodies and Banach space geometry*. Cambridge, 1999.

[33] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[34] T. Tibshirani. Principal curves revisited. *Statistics and Computing*, 2:183–190, 1992.

[35] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

[36] V. Verzelen. Data-driven neighborhood selection of a gaussian field. Technical Report 6798, INRIA, 2009.

[37] F. Villers. *Tests et sélection de modéles pour l'analyse de données protéomiques et transcriptomiques*. PhD thesis, Université Paris-Sud 11, 2007.

[38] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274, 2005.

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399